Query Expansion and Query Translation as Logical Inference

Jian-Yun Nie
Département d'informatique et recherche opérationelle
Université de Montréal
CP. 6128, succursale Centre-ville
Montreal, Quebec,
H3C 3J7 Canada
nie@iro.umontreal.ca

Abstract

A number of studies have examined the problems of query expansion in monolingual IR, and query translation for cross-language IR. However, no link has been made between them. This paper first shows that query translation is a special case of query expansion.

There is also another set of studies on inferential IR. Again, there is no relationship established with query translation or query expansion. The second claim of this paper is that logical inference is a general form that covers query expansion and query translation. This analysis provides a unified view of different sub-areas of IR.

We further develop the inferential IR approach in two particular contexts: using fuzzy logic and probability theory. The evaluation formulas obtained are shown to strongly correspond to the those used in other IR models. This indicates that inference is indeed the core of advanced IR .

1. Introduction

There have been many experiments on *Cross-Language Information Retrieval* (CLIR) during the recent years. The main problem in CLIR is query translation. Three main methods have been proposed for query translation: exploiting a bilingual dictionary, using a machine translation system or exploiting parallel texts. All the three methods are used for suggesting appropriate translation words for the original query.

Separately, there have been a number of studies on *Query Expansion* (QE). The goal of QE is to add related terms into the query so as to extend the coverage of the query, i.e. to retrieve more relevant documents. The expansion is usually done by using a thesaurus or a set of statistical association relations between terms.

Although people have observed a natural effect of query expansion during query translation in CLIR, no one has suggested a stronger relationship between them. In this paper, we show that query translation in CLIR is indeed a special case of query expansion. It is then not surprising to observe the query expansion effect in query translation.

The QE problem has been studied mostly in an ad hoc manner. A formal consideration is made from the point of view of spreading activation in a semantic network (Salton and Buckley 1988). That is, a thesaurus or a set of statistical relationships between terms is considered as a semantic network. QE is then an exploration in the semantic network so that associated terms are activated. Few have considered this process from a logical point of view. In this paper, we argue that the fundamental aspect behind query expansion is indeed *inference*. The goal of QE is to infer other forms of query through the exploitation of available knowledge (e.g. thesaurus). Therefore, QE is a means of *Inferential IR*.

Once we set up more formal relationships among CLIR, QE and logical inference, we then develop a general framework for inferential IR, which is a natural extension to inference in classical logic. Two particular developments will be considered in more detail: one in a fuzzy logic context, and another in a probabilistic context. Several previous IR models will be compared with the proposed framework. We will see that the framework corresponds well to the principles used in the previous models.

2. CLIR, query expansion and inferential IR

Let us first briefly review the common methods for CLIR, query expansion and inferential IR.

2.1. Cross-language IR

The key problem of CLIR (in addition to the common problems of IR in general) is to translate queries from a language to another. There have been mainly three approaches to this: using a machine translation (MT) system, using a bilingual dictionary, or using parallel texts.

MT systems seem to be a straightforward choice for query translation. For each query q, an MT system will give a unique translation q' for it. In some cases, the translation is reasonable. In other cases, the translation may depart from the original query. For example, the query "What effects has logging had on desertification?" is translated by Systran [1] to "Quels effets l'enregistrement a-t-il eus sur la desertification?" in French, in which "logging" has been taken in the sense of "registration". Therefore, the generated new query q' is not an equivalent to the original q. However, there is no means for us to measure the uncertainty produced during the translation of q to q' by an MT system. The proposed approach has to use q' as an equivalent to q, and considers the relationship between a document d and q' as a close approximation of that between d and q. We can see here that the translation process is separated from the search process.

On the other hand, using a bilingual dictionary, the original query is translated word by word: Each word is considered to be independent from the others in the query. We are faced with the problem of choice among multiple translations for different meanings. Several approaches have been used to cope with this problem: One can choose the first translation, assuming that it is the most common translation of the word; One can also use all the translation words together and associate to them an equal weight; Finally, the translation words can be associated with a weight according to their distribution within the document collection (e.g. the more a word appears in the collection, the higher its weight is).

Finally, by using a set of parallel texts (texts with their translations), we can extract translation relationships from them. The principle is that, the more two words (or phrases) co-occur in parallel texts (or sentences), the more one is a translation for another. This principle has been implemented in two ways. In (Yang 1998), the original query is first used to retrieve the texts in the source language from the parallel corpus that match the query. A set of keywords is then extracted from the corresponding texts in the target language. This set of keywords is used as a query translation to retrieve documents in the target language from the document collection. More often, the parallel texts are used to train a statistical translation model (Nie et al. 1999). For CLIR purposes, the IBM model 1 is usually used (Brown et al. 1992). The core of the model is a probability function $P(t|s)$ that gives the probability of translating a source word s by a target word t.

---

[1] http://babelfish.altavista.com/

2.2. Query expansion

Query expansion works as follows: Given an initial user query, some new related words are added and this forms a new query. The addition of the new words extends the original query so that it has a wider coverage than the original query. As a consequence, more relevant documents are expected to be retrieved, and the recall ratio be increased. The key problem is to identify the appropriate words to be added. Otherwise, the new query will depart from the original query in meaning. So an important question is what words should be added. Another important question is how they should be integrated into the new query.

How are new words integrated into the query?

Let us first examine this problem with respect to the two most used models: Boolean model and vector space model. In Boolean model, the added words are put into disjunction with the original query words. For example, if t is a word in the original Boolean query and t_1 is a related term to it, then t_1 is put into disjunction with t in the new query. In some cases, the added term is assigned a weight equal to that of the original term t. Thus, t is replaced by $(t \vee t_1)$. In other cases, the added term is assigned a lesser importance. So t is replaced by $(t \vee t_1^{\alpha})$ where $\alpha \leq 1$. During the evaluation, the factor α plays the role of multiplication factor. That is, if a document's similarity to t_1 is v, then its similarity to t_1^{α} is $(\alpha * v)$.

In vector space model a related word is added into the corresponding vector dimension of the query vector if it does not exist in the original vector. If it exists, its weight is increased by a certain factor. The effect of query expansion in vector space model is similar to that in Boolean model. However, the new word is not considered as an alternative (expressed as disjunction) of the original word, but as a supplement to it.

Which words to be added? - The use of thesauri or statistical associations

Automatic query expansion usually relies on a thesaurus (or pseudo-thesaurus), which stores a set of relationships between words or terms. Among the thesauri used, there are classical thesauri that are established manually (Miller 1990), or pseudo-thesauri that are established automatically using co-occurrence information (Rijsbergen 1977). In using a manual thesaurus, only strong relationships (e.g. the *is_a* relationships) are used (Voorhees 1994). In some cases, indirectly linked terms (related through more than one link) are also used, but with lower weights (Rada et al. 1991; Salton and Buckley 1988).

As to pseudo-thesauri, the co-occurrences considered are restricted within some context, which may be: document, paragraph, sentence or even some syntactic structure (Grefenstette 1992). In (Mandala et al 1999), a combined method is used: statistical analysis is used to determine the strength of relationships stored in Wordnet, according to the frequency of their co-occurrences. These strengths are then used to determine the best expansion terms (a filter) and to associate weights to them. Mandala et al. showed that when the combination takes place, the effectiveness with query expansion is much better than using Wordnet alone.

2.3. CLIR as query expansion

Although the CLIR problem has been formulated in a different way from query expansion, we can see a close relationship between them. This relationship is reflected with regard to two aspects: the principle and the knowledge used.

Principle

The query translation q' may be seen as an expansion to the original query q. The only difference lies in whether we keep the original query or not. However, this difference is not significant. If we associate a language marker with each keyword, the new query q' can be simply added to the

original query q instead of replacing it. For example, if each index is a couple <w, l> where w is a keyword and l its language marker, then the new query can be simply a mixture of words in two different languages. Similarly, documents have also to be indexed together with a language marker. In this way, query translation becomes exactly a query expansion. Therefore, we can conclude that query translation is a special case of query expansion.

It is interesting to notice that this query expansion view with language marker is an extension to the current CLIR approaches. In the current approaches, it is assumed that a document collection contains documents of the same language. The translation direction (from a language to another) is controlled manually. In reality, especially in the Internet environment, documents of different languages are mixed. By adding a language marker with indexes, these documents may be indexed together. The retrieval in different languages may be done in a single pass. The requirement is that the language of each document may be recognized automatically. This is no longer a problem as there are several automatic language identifiers (e.g. SILC[2]) that can determine the language of a text at a very high precision.

Knowledge

With respect to the knowledge used during query expansion or query translation, there is also a close relationship. In fact, a bilingual dictionary may be seen as a thesaurus: from a word, the translation relation leads to several other words in the other language. The use of parallel texts may be seen as a special case of exploiting document collection for term relationships using co-occurrences. So these two translation methods are extensions of QE techniques to a bilingual context. As to query translation by MT, there is no strict equivalent in query expansion approach. However, it is not difficult to see it as a heuristic means to derive related words from the original query.

With respect to both aspects above, we can see a close relation between query translation in CLIR and QE. We can conclude that query translation is indeed a special case of query expansion.

2.4. Inferential IR

By inferential IR, we designate all the IR approaches that try to relate a document to a query that contain different terms. In order to make such a relation, it is necessary to make inference, i.e. to infer whether a different term is related to another term (or another group of terms). Typically, if we know that there is a logical implication $t_2 \rightarrow t_1$, meaning that t_2 contains all the characteristics of t_1, then a document talking about t_2 also talks about t_1. This is one simple step in inference process. The inference process can be more complex. For example, instead to set up an implication $t_2 \rightarrow t_1$ as in classical logic, this implication may be context-dependent, i.e. it only applies when certain conditions are verified. In logic, one way to take into account the context is to place the context as an additional premise, i.e. to define $t_2 \wedge C \rightarrow t_1$ (which is equivalent to $C \rightarrow (t_2 \rightarrow t_1)$) as an implication. This implication can be understood as: t_2 implies t_1 in the context of C. In practice, C often matches part of the query, i.e. if C appears in a query, then we can conclude that t_2 in this query implies t_1.

The implication can also be uncertain. In this case, the conclusion will not be absolute as in the previous example. Instead, one would conclude that the document talks about t_2 also talks about t_1 to a certain extent (according to the degree of uncertainty of the implication).

---

[2] The SILC project, http://www-rali.iro.umontreal.ca/ProjetSILC.en.html

This is the basic principle of inferential IR. In each particular model definition, the way the inference process carries out and the way to calculate a degree of uncertainty generated by the inference process vary greatly. (Turtle and Croft 1990) suggest the use of a Bayesian network to conduct inference within a probabilistic framework. There are several recent studies on inferential IR from logical points of view (Crestani et al. 1998). In most cases, a non-classical logic framework is used because of the insufficiencies of classical logic. In this paper, however, we will limit our discussions mostly to principles.

2.5. Query expansion as inference

How is QE related to an inference process? The relation can be set up quite easily. In fact, the core of QE is twofold: the determination of appropriate expansion terms, and their integration within the query (in particular, the weighting). These are exactly the same tasks as in inferential IR, except that the framework in which query expansion is carried out may be different from that for inferential IR. If we look at the essence of both, we can see a strong relation between them by considering the relationships used in QE as implications used for inferential IR. In QE, we are interested in finding strongly related terms. Re-expressed in terms of inference, this means that we are interested in finding those terms that imply the original terms of the query. The expansion process can then be expressed as follows:

If $t_2 \rightarrow t_1$ is recognized either in a thesaurus or as a strong statistical association, then a query containing t_1 can be expanded with the term t_2.

This means that if we intend to use a relation between t_1 and t_2 in query expansion, it is equivalent to consider the relationship as a logical implication $t_2 \rightarrow t_1$, and the inference process will be equivalent to the query expansion. With this correspondence set, we can view QE as a special case of inferential IR.

In this discussion, we deliberately avoid the problem of weighting, and concentrated on the principle. In fact, as we mentioned, there are many variants in weighting. By trying to make a strong correspondence in weighting, the similarity between the principles could be less clear. However, in the next sections we will consider the weighting in more detail.

3. A general framework for inferential IR

Our general inferential framework is derived from inference in classical logic. This particular approach is taken because of two main reasons:

- We try to make the framework as simple as possible. The classical inference process is widely accepted and understood. Its use contributes to draw a simple picture of the fundamental idea of inference. The use of a simple framework also avoids the difficulty of understanding due to problems of technicality.
- The essence of most inference operations that occur in the current IR systems and models can be described by the classical inference. Once the classical inference is further enhanced with a measure of uncertainty, it can be suitable to most part of the inference process involved in current IR.

However, this is not to say that classical logic is totally satisfactory for inferential IR. One may refer to (Crestani et al. 1998) for more discussions.

Let us assume that a document d is represented as a set of terms, or equivalently as a conjunction of terms or its negation. Each term corresponds to an atom or a basic proposition. A query is a Boolean expression of terms. The relevance of a document represented by d to a query represented by q is determined by the logical implication $d \rightarrow q$. Without loss of generality, we assume that the query is in disjunctive normal form, i.e. $q = q_1 \vee q_2 \vee \ldots$, and each q_i is a

conjunction of literals. To simplify our discussion, we will use "term" to designate a literal, i.e. both a (positive) term or its negation.

A logical system is characterized by a set of logical sentences (or its closure). If we represent it by K, then the relevance of d to q with respect to this system is expressed as $K \models d \rightarrow q$. If we have $K \models d \rightarrow q$, the document is said to be relevant. If we cannot prove $K \models d \rightarrow q$, it is irrelevant.

The element K denotes *system knowledge* that makes inference possible. Notice that K in the classical Boolean model only contains tautologies. No domain-dependent knowledge is included. Therefore, $K \models d \rightarrow q$, or $d \models_K q$ is proved only if d contains all the terms required by q. For example, in the Boolean model, we will have

$$K \models (\text{computer} \wedge \text{system}) \rightarrow \text{computer}.$$

However, the system will be unable to conclude

$$K \models (\text{PC} \wedge \text{system}) \rightarrow \text{computer}.$$

because of the lack of a domain-specific knowledge PC $\rightarrow$ computer. This example shows that the standard Boolean model has little domain-specific inference capability. In order to increase its inferential power for a particular application, one has to reinforce K by adding more pieces of knowledge. By a piece of knowledge, we mean here a logical implication of the following form:

$$\text{PC} \rightarrow \text{computer}$$

$$\text{house} \rightarrow \text{home}$$

The knowledge we will consider in this paper is not limited to this classical form. In our later discussions, we will extend this form to include uncertain knowledge and contextual knowledge. For the moment, to explain the basic principle of inference in IR, let us assume this simple form of knowledge.

Once the above implications are incorporated into K of a system, the system will be able to conclude $K \models (\text{PC} \wedge \text{system}) \rightarrow \text{computer}$, as we expect it to do.

3.1. Conjunctive query

Now let us look at the inference mechanism of logic, first for a conjunctive query. We assume that K contains a set of knowledge. In classical logic, a typical inference process is made through the use of the following transitivity of implication:

$$A \rightarrow B \wedge B \rightarrow C \ \models_K A \rightarrow C$$

or

$$((A \rightarrow B \wedge B \rightarrow C) \rightarrow A \rightarrow C) \in K$$

The evaluation of $d \rightarrow q$ with the system knowledge K may be done as follows (we remove the index K because it will be always implicit):

$$d \rightarrow q' \wedge q' \rightarrow q \ \models \ d \rightarrow q$$

It means: if there is a new query q' such that the new query implies the original query, and that the new query is satisfied (implied) by a document, then we can say that the original query is also satisfied by the document.

As q' may be any query expression, we can re-write the above deduction as follows:

$$\exists q' : (d \rightarrow q' \wedge q' \rightarrow q) \ \models \ d \rightarrow q \qquad (1)$$

For example, given a query q = "house". If we know that "building", "construction" and "home" are alternative expressions of "house", then we can assume that the following implications hold:

building $\rightarrow$ house

construction $\rightarrow$ house

home $\rightarrow$ house

and of course: house $\rightarrow$ house.

If a document talks about one of these concepts, then we can also conclude that it also talks about "house". All these 4 terms are the possible expressions of *q'* in the inference (1).

This evaluation can be illustrated by the following figure:

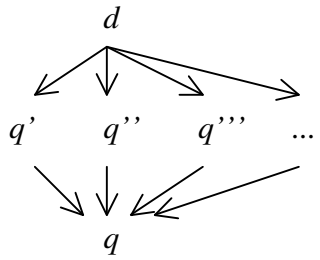

Figure 1. Illustration of inferential IR

Each path from *d* to *q* is a possible way to conclude that *d* is an answer to *q*. For each path, two conditions have to be satisfied. These conditions correspond respectively to:

- $d \rightarrow q'$: the *direct evaluation* of a query *q'* by the document *d*.
- $q' \rightarrow q$: the *relatedness* of the query *q'* to the original query *q*.

Of course, the inference is not limited in one step. One can imagine several steps: $d \rightarrow q'$, $q' \rightarrow q''$, …, $q'^{(n)} \rightarrow q$. However, we can always group all the inference steps together into $q' \rightarrow q$. In other words, $q' \rightarrow q$ can correspond to a complex inference process rather than always a simple inference step.

To estimate a degree of certainty for $d \rightarrow q$, which we denote as $P(d \rightarrow q)$, a reasonable equation is as follows:

$$P(d \rightarrow q) = P(\bigvee_{q'} (d \rightarrow q' \wedge q' \rightarrow q)) \qquad (3)$$

In a symmetric way, we can also consider alternative document expressions instead. The expression equivalent to (3) would be:

$$P(d \rightarrow q) = P(\bigvee_{d'} (d \rightarrow d' \wedge d' \rightarrow q)) \qquad (3')$$

In this expression the inferential process is included in the $d \rightarrow d'$ whereas $d' \rightarrow q$ corresponds to a direct query evaluation without inference. The expressions (3) and (3') correspond respectively to aquery-driven and document-driven approach.

The expression (3') corresponds exactly to the uncertainty principle expressed by van Rijsbergen (Rijsbergen 1986): The uncertainty of $d \rightarrow q$ is determined by the minimal extent to which we have to change the expression of *d* to *d'* (i.e. $d \rightarrow d'$) such that *q* becomes satisfied in the changed *d'* (i.e. $d' \rightarrow q$). In this paper, we will focus on the query-driven approach.

3.2. Disjunctive query

A disjunctive query ($q_1 \vee q_2$) expresses two alternatives, either q_1 or q_2 has to be satisfied. This is not different from the alternative q's we have considered for the evaluation of conjunctive queries. Therefore, the same evaluation method can be applied:

$$P(d \rightarrow (q_1 \vee q_2)) = P(d \rightarrow q_1 \vee d \rightarrow q_2) \quad (4)$$

This decomposition may be viewed as an additional step of finding alternative query expressions, as shown in Figure 2(a). All the alternatives to each disjunct can also be viewed directly as alternatives to the whole original query, as shown in Figure 2(b). In both cases, its evaluation does not represent any additional difficulty.

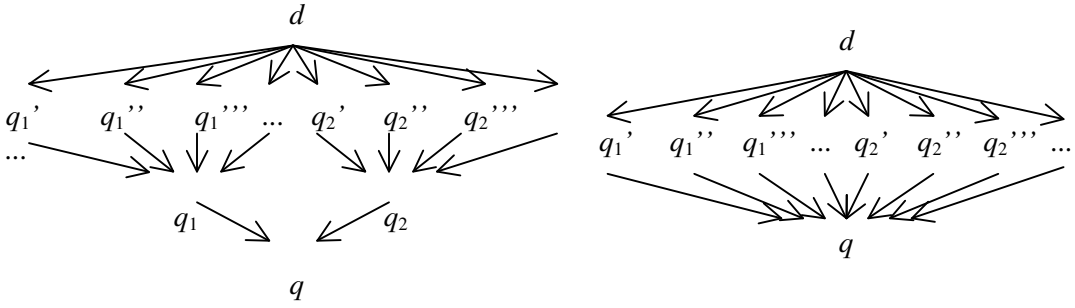

Figure 2. Evaluation of a disjunctive query

In what follows, we will focus on the evaluation of conjunctive queries. Our aim is to further develop the right side of Equation (3). For this, we will have to consider particular settings for the evaluation of conjunction and disjunction. In this paper, we examine two different settings, one in fuzzy logic, and another in probabilistic framework.

4. Fuzzy logic setting

A particular setting to further develop the equation (3) is important to determine how the logical operators $\wedge$ and $\vee$ can be evaluated. Fuzzy logic (or fuzzy set theory) provides a relatively easy way to deal for this. In fuzzy logic evaluations, there is no particular consideration on the dependency or independency between the elements connected by these operators. The evaluations are completely compositional. For example, the numerical operators that correspond to these logical operators are often min and max.

$$P(A \wedge B) = \min(P(A), P(B))$$

$$P(A \vee B) = \max(P(A), P(B)).$$

No dependency between the connected elements is considered. This account makes it easy to develop of the right side of Equation (3). If we use min and max for conjunction and disjunction, then:

$$\begin{aligned} P(d \rightarrow q) &= P(\bigvee_{q'} (d \rightarrow q' \wedge q' \rightarrow q)) \\ &= \max_{q'} (P(d \rightarrow q' \wedge q' \rightarrow q)) \\ &= \max_{q'} (\min(P(d \rightarrow q'), (q' \rightarrow q))) \end{aligned}$$

Although min and max are most often used in studies in fuzzy set theory, they may be inappropriate for our task at hand. The main objection may be the fact that in min and max, the weakest and the strongest element predominates completely. This may be counterintuitive particularly for the combination $d \rightarrow q'$ and $q' \rightarrow q$. For example, given two possible q's: q'_1 and q'_2; if $P(d \rightarrow q'_1) = 0.1$, $P(q'_1 \rightarrow q) = 0.1$, and $P(d \rightarrow q'_2) = 0.1$, $P(q'_2 \rightarrow q) = 1.0$, both paths will lead to the same evaluation of $P(d \rightarrow q) = 0.1$. Intuitively, however, the path through q'_2 seem much more certain than that through q'_1.

It seems then more intuitive to use an evaluation of conjunction in which both elements contribute simultaneously. One of such fuzzy evaluations is as follows:

$$P(A \wedge B) = P(A) * P(B)$$

$$P(A \vee B) = P(A) + P(B) - P(A) * P(B).$$

Let us assume two fuzzy functions Δ and ∇ respectively for the evaluations of conjunction ($\wedge$) and disjunction ($\vee$). We further assume the following properties on them (where x, y and z are fuzzy values)[3]:

$$\Delta(x, y) = 1 - \nabla(1-x, 1-y)$$

$$\Delta(x, x) = x; \qquad \text{(idempotence)}$$

$$\Delta(x, y) = \Delta(y, x);$$

$$\Delta(x, \Delta(y, z)) = \Delta(\Delta(x, y), z) ;$$

$$\Delta(x, \nabla(y, z)) = \nabla(\Delta(x, y), \Delta(x, z)) \qquad \text{(distributivity)}$$

Then formula (3) can be developed as follows:

$$P(d \rightarrow q) = P(\bigvee_{q'} (d \rightarrow q' \wedge q' \rightarrow q))$$

$$= \nabla_{q'}[\Delta(P(d \rightarrow q'), P(q' \rightarrow q))]$$

Let us abbreviate the direct evaluation $P(d \rightarrow q')$ as $P_d(q')$, i.e. the degree of certainty that d provides a direct answer to q'. This value may be the one provided by a classical IR evaluation. Let us further use the following *weighted query* A^{β}, and define its evaluation $P_d(q^{\beta})$ as $\Delta(P_d(A), \beta)$. The weighted query is used to express an alternative query form that has a β-relatedness to the original query q. Using the weighted query, equation (4) can be expressed as follows:

$$P(d \rightarrow q) = \nabla_{q'} [P_d(q'^{P(q' \rightarrow q)})]$$

$$= P_d(\bigvee_{q'} q'^{P(q' \rightarrow q)}) \qquad (5)$$

The expression

$$Exp(q) = \bigvee_{q'} q'^{P(q' \rightarrow q)} \qquad (6)$$

expresses exactly the essence of query expansion in Boolean model: A query is expanded into a disjunction of all the alternative forms. Each alternative form is associated with its relationship with the original query. The evaluation $P(d \rightarrow q)$ is then determined by $P_d(Exp(q))$.

For example, if the original query is "house", and we have the same related words as before, then the expanded query $Exp(q)$ is

---

[3] We will see later that no existing fuzzy operators have all these properties.

$$\text{house}^{1} \vee \text{building}^{P(\text{building} \rightarrow \text{house})} \vee \text{construction}^{P(\text{construction} \rightarrow \text{house})} \vee \text{home}^{P(\text{home} \rightarrow \text{house})}$$

This is exactly what we obtain with the query expansion method in Boolean model.

Now let us consider a more complex form of query: $(q_1 \wedge q_2)$.

$$P(d \rightarrow (q_1 \wedge q_2)) = P_d[\bigvee_{q'}(q'^{P(q' \rightarrow q_1 \wedge q_2)})]$$

$$= P_d[\bigvee_{q'}(q'^{P(q' \rightarrow q_1 \wedge q' \rightarrow q_2)})]$$

Assume that an alternative expression q' is also of the form $(q'_1 \wedge q'_2)$ and that q'_1 is related to q_1 and q'_2 to q_2. Then

$$P(d \rightarrow (q_1 \wedge q_2)) = P_d[\bigvee_{q'1,\, q'2}(q'_1{}^{P((q'_1 \wedge q'_2) \rightarrow q_1 \wedge q_2)})]$$

$$= P_d[\bigvee_{q'1,\, q'2}(q'_1{}^{P((q'_1 \wedge q'_2) \rightarrow q_1)} \wedge q'_2{}^{P((q'_1 \wedge q'_2) \rightarrow q_2)})] \qquad (4)$$

It is not possible to further decompose the expression on the right side. A further decomposition is possible with the following independence assumption.

Independence assumption

We assume that in the above evaluation, $(q'_1 \wedge q'_2) \rightarrow q_1$ is determined solely by $q'_1 \rightarrow q_1$ and $(q'_1 \wedge q'_2) \rightarrow q_2$ solely by $q'_2 \rightarrow q_2$.

With the independence assumption, the above equation becomes:

$$P(d \rightarrow (q_1 \wedge q_2)) = P_d[\bigvee_{q'_1,\, q'_2}(q'_1{}^{P(q'_1 \rightarrow q_1)} \wedge q'_2{}^{P(q'_2 \rightarrow q_2)})]$$

$$= P_d[\bigvee_{q'_1}(q'_1{}^{P(q'_1 \rightarrow q_1)}) \wedge \bigvee_{q'_2}(q'_2{}^{P(q'_2 \rightarrow q_2)})]$$

That is:

$$Exp(q_1 \wedge q_2) = \bigvee_{q'_1}(q'_1{}^{P(q'_1 \rightarrow q_1)}) \wedge \bigvee_{q'_2}(q'_2{}^{P(q'_2 \rightarrow q_2)}) \qquad (4')$$

In other words, the expansion of a compound query $(q_1 \wedge q_2)$ may be done by expand q_1 and q_2 separately. For example, if our query is q=(house $\wedge$ garden), and we assume that "yard" is a related word to "garden", then the expanded query would be:

$$[\text{house}^{1} \vee \text{building}^{P(\text{building} \rightarrow \text{house})} \vee \text{construction}^{P(\text{construction} \rightarrow \text{house})} \vee \text{home}^{P(\text{home} \rightarrow \text{house})}]$$

$$\wedge \quad [\text{garden}^{1} \vee \text{yard}^{P(\text{yard} \rightarrow \text{house})}]$$

This corresponds exactly to the common query expansion process in Boolean model. This shows that QE can be derived from the general inference framework.

4.1. The problem of appropriate fuzzy operators

We have assumed a set of properties for Δ and ∇. In reality, no dual fuzzy operators have all these properties. The closest operators are *triangular norm* and *co-norm* (Dubois and Prade 1984). A triangular norm Δ is used for the evaluation of conjunction, while its *co-norm* is used for the evaluation of disjunction.

A triangular norm Δ is a function $[0,1] \times [0,1] \rightarrow [0,1]$ that verifies the following conditions (where x, x', y, y', z are fuzzy values in [0,1]):

1. $\Delta(x, y) = \Delta(y, x)$;
2. $\Delta(x, \Delta(y, z)) = \Delta(\Delta(x, y), z)$

3. If $x \geq x'$, and $y \geq y'$, then $\Delta(x, y) \geq \Delta(x', y')$.

The function *min* is a triangular norm. Its co-norm is *max*. Multiplication of real numbers (*) is another triangular norm. Its co-norm is $(x + y - x*y)$. These two sets of functions are among the most used functions for logical operators in fuzzy set theory.

However, these operators do not have the distributivity. The second couple of norm and co-norm also do not verify idempotence. This means that if a norm and a co-norm are used, we cannot have all the logical properties we desire. Some properties have to be dropped. This implies an approximation in the evaluation process.

The problem comes from the fact that no relationship is considered between the disjuncts and the conjuncts when they are combined. In the min-max case, the two conjuncts *A* and *B* are considered to be strongly related - one of them entails the other. This may be viewed as in one of the following situations:

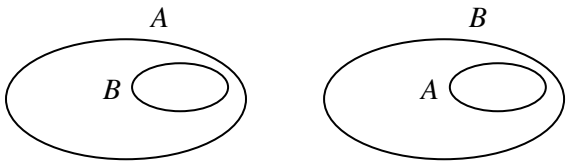

Figure 3. The relationships assumed by the min-max evaluations.

In the case of multiplication as norm, the two elements are considered to be independent (one can compare the evaluation formulas with those in probability theory for independent events). Such a uniform view for any pair of combined elements is certainty not always reasonable. It represents a strong simplification. However, by this simplification, we gain the simplicity of decomposing complex expressions into simpler ones, i.e. a stronger compositionality in the evaluations.

4.3. Discussion

In our development, we have made use of the compositionality on the evaluation of $A \wedge B$, and the independent assumption. For the (home $\wedge$ garden) example, the compositionality allows us to transform:

(home $\wedge$ yard) $\rightarrow$ (house $\wedge$ garden)

into:

(home $\wedge$ yard $\rightarrow$ house) $\wedge$ (home $\wedge$ yard $\rightarrow$ garden)

The independence assumption then transforms it into:

(home $\rightarrow$ house) $\wedge$ (yard $\rightarrow$ garden)

These two implications are then evaluated separately because of the compositionality of the fuzzy operations. This assumption is not always reasonable. For example, while we can consider the relationships (company $\rightarrow$ house) and (park $\rightarrow$ garden) as being strong in general, the relationship (company $\wedge$ park) $\rightarrow$ (house $\wedge$ garden) is much weaker. In fact, (house $\wedge$ garden) cannot be decomposed and evaluated separately. One element serves as a context to another. To determine whether (company $\rightarrow$ house) and (park $\rightarrow$ garden) hold in this case, one has to use the

contextual information. What we need to determine is $P((\text{company} \rightarrow \text{house}) \mid \text{yard})$ and $P((\text{park} \rightarrow \text{garden}) \mid \text{house})$, i.e. the relationships in the context of "yard" and "house" respectively[4].

Notice that the above contextual expression is not associated with a strict meaning in fuzzy logic (or in logic in general). In fact, no special operator in fuzzy logic has been defined in fuzzy logic to take into account such a contextual aspect. It is neither the goal of the present paper to develop such a "standard" method. However, we can state that this contextual information is important in IR. This has been demonstrated in several experiments. For example, (Qiu and Frei 93) showed that a higher effectiveness can obtain if query expansion considers the relationships of expansion terms with the whole query rather than the relationships with one single term. In our terms, this means that it is better to consider $P((\text{company} \rightarrow \text{house}) \mid \text{yard})$ than $P(\text{company} \rightarrow \text{house})$.

A similar method has been used for query translation (e.g. (Gao et al. 2000)): in order to determine the best translation word of a query, not only the relationship between a possible translation word with one of the original words has to be considered, but also the relationship with other words. The other words together form a context for the selection of the translation word of one particular word of the query.

This same approach has been used in query expansion in (Mandala et al.1999): the expansion term is determined according to the relationship of the candidate expansion term to all the original terms inn the query.

The problem of weighting of expansion terms is an important issue. In (Voorhees 1993) and (Voorhees 1994), expansion terms found in Wordnet are added into the query vector with a uniform weighting. This expansion brings rather a decrease in retrieval effectiveness. In (Mandala et al. 1999), the same relationships are used, however with a more appropriate weighting that is calculated according to co-occurrences. In the experiments of Mandala et al., the use of Wordnet brings a significant improvement in effectiveness.

5. Probabilistic setting

Now we will develop the right side of Equation (3) in a probabilistic framework. Let us denote by $d \rightarrow_{q'} q$ the evaluation of $d \rightarrow q$ with one single path through q', i.e.:

$$d \rightarrow_{q'} q = d \rightarrow q' \wedge q' \rightarrow q.$$

Equation (3) can the be rewritten as follows:

$$P(d \rightarrow q) = P(\bigvee_{q'} (d \rightarrow_{q'} q))$$

In the fuzzy logic setting, we assumed that each evaluation path is independent fro the others. We pointed out that this assumption is not reasonable. Let first fist examine a disjunction of two elements: $A_1 = d \rightarrow_{q_1} q$ and $A_2 = d \rightarrow_{q_2} q$. Without the independent assumption between A and B we have:

$$P(A_1 \vee A_2) = P(A_1) + P(A_2) - P(A_1)*P(A_2 \mid A_1) = P(A_1) + P(A_2) - P(A_2)*P(A_1 \mid A_2)$$
$$= P(A_1) + P(A_2) - (1/2)\,[P(A_1)*P(A_2 \mid A_1) + P(A_2)*P(A_1 \mid A_2)]$$

[4] In classical logic, this dependencies on context are rather expressed as yard $\rightarrow$ (company $\rightarrow$ house) and house $\rightarrow$ (park $\rightarrow$ garden). However, this expression may puzzle some readers. Therefore, we use an expression closer to the more familiar notation in probability theory.

Notice that an average is used in the last expression because we anticipate the approximations that will occur in probability estimations. The use of the average brings more robustness to errors in probability estimates.

We can also reasonably assume that:

$$P(A_2 \mid A_1) = P(d \rightarrow_{q_2} q \mid d \rightarrow_{q_1}) = P(q_2 \mid q_1)$$

Therefore,

$$P(A_1 \vee A_2) = P(A_1) + P(A_2) - (1/2)\,[P(A_1) * P(q_2 \mid q_1) + P(A_2) * P(q_1 \mid q_2)]$$

For a disjunction of more elements $\bigvee_{i=1}^{n} A_i = A_1 \vee A_2 \vee A_3 \vee \ldots$, the expression is more complex:

$$P(\bigvee_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) - (1/2) \sum_{i=1}^{n} \sum_{j=1}^{n} P(A_i) * P(q_j \mid q_i) + \ldots$$

As in other probabilistic models, this formula is too complex to be calculated in practice. To simplify the formula, we will limit to the consideration of pairwise-dependencies, i.e. we only consider $P(d \rightarrow_{q_j} q \mid d \rightarrow_{q_i} q)$ and assume that higher-order dependencies are null. This assumption is similar to the binary-dependence assumption made in (Rijsbergen 1979).

Therefore,

$$P(d \rightarrow q) = \sum_{i=1}^{n} P(A_i) - (1/2) \sum_{i=1}^{n} [P(A_i) \times \sum_{j=1}^{n} P(q_j \mid q_i)] \qquad (5)$$

Now let us consider in more detail the evaluation of one path $P(A_i) = P(d \rightarrow_{q_i} q) = P(d \rightarrow q_i \wedge q_i \rightarrow q)$. It may be reasonably assume that $d \rightarrow q_i$ and $q_i \rightarrow q$ are independent, i.e. the direct evaluation of q_i with respect to d is independent from the relationship of q_i to the original query. Therefore,

$$P(d \rightarrow q_i \wedge q_i \rightarrow q) = P(d \rightarrow q_i) * P(q_i \rightarrow q).$$

Let us use the conditional probabilities $P(q_i \mid d)$ and $P(q \mid q_i)$ to evaluated respectively $P(d \rightarrow q_i)$ and $P(q_i \rightarrow q)$. $P(q_i \mid d)$ is the direct evaluation of q_i in, which can be estimated as in other probabilistic models. For $P(q \mid q_i)$ we have:

$$P(q \mid q_i) = \frac{P(q_i \mid q) * P(q)}{P(q_i)}$$

Because $P(q)$ is a constant, we have:

$$P(q \mid q_i) \propto \frac{P(q_i \mid q)}{P(q_i)}$$

and $$P(A_i) \propto P(q_i \mid d) \times \frac{P(q_i \mid q)}{P(q_i)}.$$

Putting them into Equation (5), we obtain,

$$P(d \rightarrow q) \propto \sum_{i=1}^{n} P(q_i \mid d) \times \frac{P(q_i \mid q)}{P(q_i)} - (1/2) \sum_{i=1}^{n} [P(q_i \mid d) \times \frac{P(q_i \mid q)}{P(q_i)} \times \sum_{j=1}^{n} P(q_j \mid q_i)] \qquad (6)$$

5.1. Discussions

Equation (6) may seem different from previous formulations of probabilistic models. In fact, the main elements contained in it can be compared with those used in other models.

- $P(q_i \mid d)$: This conditional probability determines the direct satisfaction of an alternative query q' by the document d. It can be estimated just as in the other probabilistic models.
- $1/P(q_i)$: This is a factor comparable to the common IDF factor. The more a query q_i is composed of common terms, the higher is $P(q_i)$, and then the lesser is this factor.

The above two factors are commonly used in other models. However, the difference is, in other models, the above relationships are estimated with the original query q, whereas we use an alternative query q_i instead. By this alternative query, we can represent either the original query (i.e. q_i is identical to q) or a query that has been changed through an inference process. Once a change is made we have also take it into account. This is the role of he following factors:

- $P(q_i \mid q)$: This probability denotes the strength to which an alternative query q_i is related to the original query. This element is exactly an evaluation of the inference process. Notice that the original query q is a particular case of q_i. When $q_i = q$, we have $P(q_i \mid q) = 1$.
- The last element $\sum_{i=1}^{n}[P(q_i \mid d) \times \frac{P(q_i \mid q)}{P(q_i)} \times \sum_{j=1}^{n} P(q_j \mid q_i)]$ is added in order to take into account the relationships between different inference paths. This element is important in the case where no particular constraints are made on the inference paths. One can even use the same path twice. It is then important to determine whether two paths are independent or not.

If no inference takes place, the above formula reduces to the following form, which is more comparable to the previous probabilistic models:

$$P(d \rightarrow q) \propto \frac{P(q \mid d)}{P(q)} \propto P(q \mid d).$$

The inference process, however, increases the reasoning capability of a model. Without this process, a model can only used the features (terms) directly observed in a document to compare to those stated in a query, and only identical features are compared. The addition of the inference process allows the model to be able to compare two sets of different features (respectively for a document and a query). If there are ways to establish a relationship between two different features, then we can still relate the document to the query.

5.2. Independent model

We consider that a query q is a conjunction of terms, and we assume that these terms are mutually independent as in the binary independent model (Rijsbergen 1979). As in the development wit fuzzy setting, we assume further that a query is expanded term-by-term. Therefore, q_i contains as many terms as in the original query q. Assuming that q contains m terms, we will have:

$$P(q_i \mid d) = \prod_{j=1}^{m} P(t_{ij} \mid d)$$

$$P(q_i) = \prod_{j=1}^{m} P(t_{ij})$$

and $$P(q_i \mid q) = \prod_{j=1}^{m} P(t_{ij} \mid q)$$

As in the fuzzy logic setting, we can further assume that an expansion term is selected according to its relationship with one term in the original query, and that the other terms in the query do not have any impact on it. This assumption leads to the following simplified form:

$$P(q_i \mid q) = \prod_{j=1}^{m} P(t_{ij} \mid t_i) \qquad (6)$$

Although this last simplification may make the calculation simpler (this simplification has often been made in previous experiments), it is not always reasonable. Recall the (house $\wedge$ garden) example we showed in the fuzzy setting. If we select related terms for each original term "house" and "garden" separately, we can end up with a strong candidate for alternative query that is (company $\wedge$ park), because both P(company | house) and P(park | garden) may be strong. However, such an alternative query implies a strong derivation in meaning from the original one. This candidate can be eliminated if we consider P(company | house $\wedge$ garden) and P(park | house $\wedge$ garden) instead. In particular, the probability P(company | house $\wedge$ garden) would be much weaker than P(company | house).

5.3. Comparison with some previous models

The above formulation can be compared (at least in principles) to some previous probabilistic models in which one tries to integrate some inferential power. We can first compare it with the binary dependent probabilistic model (Rijsbergen 1979). This model takes into account the dependencies between pairs of terms, i.e. $P(t_{ij} \mid t_i)$. We can see that this is exactly the independent model with the simplification (6). Nevertheless, in comparison with the independent model, this model is able to incorporate some limited relationships between terms, and the use of such relationships in the retrieval process precisely corresponds to a form of inference.

In (Fuhr 1991), two classical probabilistic models are called independent indexing model and independent retrieval model. In the first model, the features (terms) included in each document that is manually judged are used to determine the types of query that this document is able to satisfy. This is a generalization on query. On the other hand, the second model tries to determine the characteristic features of a particular query from a set of manual judgements. A generalization is made on documents, i.e. if a document contains features that correspond to those for relevant documents for that query, then the document has a high probability to be relevant. As Fuhr noticed, these models cannot generalize a set of manual judgments on both document and query. The new model suggested by Fuhr tries to make such a two-directional generalization. Although no particular formulation has been suggested in (Fuhr 1991), it is can be seen that the key is to establish relationships between different features. This principle is similar to the inference process we propose in this paper.

The Bayesian network model proposed in (Turtle and Croft 90) is a probabilistic model that incorporates explicitly an inference step. Inference is made to determine the relationship between a document and a query through a Bayesian network. This network (Figure 4) connects the document to a set of independent indexes (together with probability estimates), which is then connected to a set of independent terms used by the users. Finally, the terms used by the users are related to a particular query.

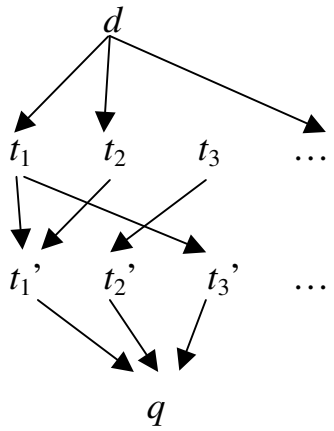

Figure 4. A fragment of Bayesian network

The additional inference power is mainly generated due to the connection between t_i and t_j, which are respectively included in the document and the query. It is these connections that allow the system to retrieve (in principle) a document that contain terms different from but related to those in the query. Each connection between the two term layers is associated with a matrix of probabilities. For example, the two connections from t_1 and t_2 to t_1' is associated with the following matrix:

$$\begin{vmatrix} P(\neg t_1' | \neg t_1, \neg t_2) & P(t_1' | \neg t_1, \neg t_2) \\ P(\neg t_1' | t_1, \neg t_2) & P(t_1' | t_1, \neg t_2) \\ P(\neg t_1' | \neg t_1, t_2) & P(t_1' | \neg t_1, t_2) \\ P(\neg t_1' | t_1, t_2) & P(t_1' | t_1, t_2) \end{vmatrix}$$

Once the combinations of absence and presence of (t_1, t_2) are activated to some degree by the document, t_1' and $\neg t_1$' will be activated in turn to some degree, and this contributes to determine the probability of q (through another matrix of probabilities). Notice that this approach corresponds more to the document-driven approach we mentioned earlier (Equation (3')). In this case, t_1'is considered to be a related term to the original document expression d, part of which being combinations of (t_1, t_2). If this Bayesian network is reversed, and connections are made from query to document, then the relationships encoded by the connections between the two layers of terms correspond exactly to what is described in this paper.

Notice further that the connections between terms in the Bayesian network are not restricted to pairwise connections, i.e. the simplification (6) is not assumed.

These comparisons show that the general framework proposed in this paper correspond well to the current approaches suggested in the advanced IR models.

6. Conclusion

In the last sections, we have developed two possible directions for inferential IR, one within fuzzy logic setting and the other with a probabilistic setting. The resulting evaluation formulas can be compared with those proposed in the previous IR models. This comparison provides some validation indications that our developments correspond well to the common practice in IR. This leads to the main claim in this paper: inference is the core part of advanced IR. Although it is not expressed as such, many modern IR systems do include more or less inferential power. The goal of this paper is to make this core part stand out more distinctly.

We took the classical inference as the starting point due to the following reasons:

- The classical inference process is widely accepted and understood. Its use contributes to draw a simple picture of the fundamental idea of inference.
- Once the classical inference is enhanced with a measure of uncertainty, it can be suitable to most part of the inference process involved in current IR.

This does not mean that the classical logic is necessarily a sufficient framework. To see the insufficiencies, one can refer to (Crestani et al. 1998).

We also examined two particular sub-areas in IR: cross-language IR and query expansion. Despite the fact that the knowledge used in these approaches may be very different, we can still consider query translation as a particular case of query expansion, and query expansion as a particular case of inference. This analysis provides a unified view of different sub-areas.

During the two developments of the general inference framework, we have made a series of simplification assumptions in order to correspond to the current IR models. These assumptions, however, are not all reasonable. Several questions may be raised: Are they a source of problem in query expansion and CLIR that did not surface until now? Can we explain some failures in query expansion by these simplifications? These are the questions we need to examine.

References

P. F. Brown, S.A.D. Pietra, V. D. J. Pietra, and R. L. Mercer (1992). The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19: 263-312.

F. Crestani, M. Lalmas and C.J. van Rijsbergen (eds.) (1998). *Information Retrieval, Uncertainty and Logics*, Kluwer Academic Publishers, pp. 17-38.

D. Dubois and H. Prade (1984). Fuzzy logics and the generalized modus ponens revisited. *Cybernetics and Systems: An International Journal*, 15: 293-331.

N. Fuhr (1991) Probabilistic models in information retrieval, *The Computer Journal*, 35 (3): 243-255.

J. Gao, J.Y. Nie, E. Xun, J. Zhang, M. Zhou, C. Huang (2001) Improving Query Translation for CLIR using Statistical Models, *ACM-SIGIR'01*, New Orleans, pp. 96-104.

G. Grefenstette (1992). Use of syntactic context to produce term association lists. *ACM-SIGIR'92*, 89-97.

R. Mandala, T. Tokunaga, H. Tanaka (1999). Combining multiple evidence from different types of thesaurus for query expansion, *ACM-SIGIR'99,* pp. 191-197.

G. Miller (ed.) (1990). *Wordnet: an on-line lexical database*, *International Journal of Lexicography.*

J.-Y. Nie, M. Simard, P. Isabelle and R. Durand (1999). Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web, *ACM-SIGIR'99*, Berkeley, pp. 74-81.

Y. Qiu and H. P. Frei (1993). Concept based query expansion. *Research and Development in Information Retrieval, ACM-SIGIR'93*, 160-169.

R. Rada, J. Barlow, J. Potharst, P. Zanstra, and D. Bijstra (1991). Document ranking using an enriched thesaurus. *Journal of Documentation*, 47: 240-253.

C. J. van Rijsbergen (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33: 106-119.

C. J. van Rijsbergen (1979). *Information Retrieval*, 2nd ed. Butterworths: London.

C. J. van Rijsbergen (1986). A non-classical logic for information retrieval. *The Computer Journal*, 29(6): 481-485.

G. Salton and C. Buckley (1988). On the use of spreading activation methods in automatic information retrieval. *ACM-SIGIR'88*. pp. 147-160.

H. Turtle and W. B. Croft (1990). Inference network for document retrieval. *ACM-SIGIR'90*, pp. 1-24.

Y. Yang, J.G. Carbonell, R.D. Brown, R.E. Frederking (1998). Translingual information retrieval: learning from bilingual corpora, *Artificial Intelligence*, 103: 323-345.

E. M. Voorhees (1993). Using WordNet to disambiguate word senses for text retrieval. *ACM-SIGIR'93*, pp. 171-180.

E. M. Voorhees (1994). Query expansion using lexical-semantic relations. *ACM-SIGIR'94*, Dublin, 61-70.